# Modèle des blocs latents avec une classe de bruit

Vincent Brault, Charlotte Laclau

# Modèle des blocs latents avec une classe de bruit

Vincent Brault [1] & Charlotte Laclau [2]

[1] *Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France*
*vincent.brault@univ-grenoble-alpes.fr*
[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
*charlotte.laclau@univ-grenoble-alpes.fr*

**Résumé.**

La classification croisée est connue pour être une approche très efficace en apprentissage non supervisé en raison de sa capacité à partitionner simultanément les lignes et colonnes d'une matrice de données. Cependant, dans un contexte de grande dimension, les méthodes de classification croisée peuvent être perturbées en raison de la présence de colonnes bruitées et/ou non discriminantes. Dans cet exposé, nous abordons ce problème en proposant un nouveau modèle de classification croisée, à partir du modèle des blocs latents, qui modélise l'existence d'une classe de bruit, à laquelle appartient l'ensemble de ces variables non pertinentes pour le partitionnement des données. Les résultats obtenus sur des données synthétiques montrent l'efficacité de notre modèle dans le contexte des données bruitées en grande dimension. Enfin, nous soulignons l'intérêt de cette approche sur deux jeux de données réelles initialement proposés pour étudier les diversités génétiques à travers le monde.

**Mots-clés.** Classification croisée, apprentissage statistique, modèle des blocs latents, sélection de variables

**Abstract.**

Co-clustering is known to be a very powerful and efficient approach in unsupervised learning because of its ability to partition data based on both modes of a dataset. However, in high-dimensional context co-clustering methods may fail to provide a meaningful result due to the presence of noisy and/or irrelevant features. In this talk, we propose to tackle this issue by proposing a novel co-clustering model, based on the latent block model, and which assumes the existence of a noise cluster, that contains all irrelevant features. Experimental results on synthetic datasets show the efficiency of our model in the context of high-dimensional noisy data. Finally, we highlight the interest of the approach on two real datasets which goal is to study genetic diversity across the world.

**Keywords.** Co-clustering, Statistical Learning, Latent Blocks Model, Feature Selection

# 1 Introduction

L'apprentissage statistique est devenu un outil vital pour l'analyse efficace d'un grand volume de données qui a trouvé son application dans de nombreux domaines de notre vie quotidienne tels que le marketing, la reconnaissance d'objets ou encore la génétique. Traditionnellement, la plupart des techniques développées pour l'analyse de ces grands volumes de données reposent sur le paradigme de l'apprentissage supervisé dont l'objectif est de construire une fonction de décision à partir d'une base d'observations auxquelles sont associées des étiquettes de sorties. Cependant, dans de nombreuses situations, l'étiquetage des données peut s'avérer impossible en raison par exemple du volume trop important des données ou du coût d'acquisition des étiquettes qui peut être trop élevé. Dans le cas où l'on dispose de données non étiquetées, on parle alors d'apprentissage non supervisé. Parmi les différentes approches non supervisées existantes, la classification croisée, ou *co-clustering*, implique un regroupement simultané des observations (lignes) et des variables (colonnes) d'une matrice de données, de manière à faire apparaître des blocs homogènes. Or, dans de nombreux cas, et notamment dans le cas des données de grande dimension, une proportion significative des colonnes de la matrice ne fournissent aucune information et ne possèdent aucune structure particulière. Dans cette situation, tenter d'apprendre en incluant cette partie des données, que l'on peut qualifier de bruits, perturbe fortement les algorithmes de classification croisée et peuvent masquer la structure des données. Typiquement, les algorithmes vont parfois se sentir obligés de subdiviser des classes qui ne sont pas discriminantes pour essayer d'obtenir de l'homogénéité. Dans ce travail, notre objectif est de partir de l'hypothèse qu'une partie des variables n'apportent aucune information pertinente et devraient par conséquent être écartées du partitionnement des données. Pour y parvenir, nous nous intéressons plus particulièrement à un modèle probabiliste, le modèle des bloc latent (voir [3]), et supposons qu'il existe une classe de bruit pour les colonnes de la matrice. Les variables appartenant à cette classe sont supposées être tirées selon une distribution de probabilité qui ne dépend pas de la structure en blocs, et par conséquent n'impacteront pas la classification des lignes de la matrice.

Nous commençons par présenter le modèle proposé en détaillant les différences avec le modèle des blocs latents. Dans un deuxième temps, nous présentons l'intérêt de notre approche sur des données réelles pour lesquelles l'objectif est d'étudier la diversité génétique à travers le monde.

# 2 Modèle

Le modèle des blocs latents avec une classe de bruit étant une extension du modèle des blocs latents, nous commençons par présenter ce premier afin de permettre une meilleure compréhension des différences.

## 2.1 Modèle des blocs latents

Soit $\mathbf{x} = (x_{ij})_{i=1,\dots,n;j=1,\dots,d} \in \{0,1\}^{n\times d}$ une matrice de données binaires de dimension $n \times d$ mettant en relation $n$ objets (observations) et $d$ variables (attributs). L'objectif du modèle des blocs latents est d'opérer des permutations sur les lignes et sur les colonnes pour obtenir une réorganisation faisant apparaître des blocs contrastés. La partition $\mathbf{z}$ d'un échantillon $\{1,\dots,n\}$ en $g$ classes est représentée par la matrice de classification $(z_{i,k})_{i=1,\dots,n;k=1,\dots,g}$ où $z_{ik} = 1$ si $i$ appartient à la classe $k$ et 0 sinon ; parfois, nous utiliserons la notation duale $z_i = k$. De façon similaire, la partition $\mathbf{w}$ d'un échantillon $\{1,\dots,d\}$ en $m$ classes est représentée par la matrice de classification $(w_{j\ell})_{j=1,\dots,d;\ell=1,\dots,m}$. Les variables aléatoires sont notées en majuscule, une colonne $j$ d'une matrice $(a_{ij})$ est représentée par $a_{\cdot j} = (a_{1j},\dots,a_{nj})^T$ et la somme de toutes ses cases est notée $a_{+j}$.

Le modèle des blocs latents repose sur trois hypothèses :

1. Les distributions des variables latentes sont indépendantes $p(\mathbf{Z},\mathbf{W}) = p(\mathbf{Z})p(\mathbf{W})$.
2. L'affection $z_i$ de chaque ligne $i$ à une classe est indépendante des autres affectations et suit une multinomiale $\mathcal{M}(1;\pi_1,\dots,\pi_g)$ ($\pi_k$ est donc la probabilité pour une ligne d'appartenir à la classe $k$). De même, $\tau_\ell$ est la probabilité d'appartenance d'une colonne à la classe $\ell$.
3. Connaissant le couple de partitions $(\mathbf{z},\mathbf{w})$, les variables $X_{ij}$ sont indépendantes et issues d'une loi de Bernoulli dont le paramètre $\alpha_{k\ell}$ ne dépend que du bloc dans lequel elles se trouvent :

$$X_{ij}|z_{ik} = 1,\, w_{j\ell} = 1 \sim \mathcal{B}(\alpha_{k\ell}).$$

## 2.2 Modèle des blocs latents avec une classe de bruit

Dans le modèle des blocs latents avec une classe de bruit, nous posons deux nouvelles hypothèses de modélisation : (1) la matrice de données observées est générée selon un produit de deux mélanges de fonctions de densité de probabilités sous-jacentes où le premier mélange est associé à une structure de blocs pertinente tandis que le second ne contient que les caractéristiques non pertinentes de la structure ; (2) par conséquent, nous modélisons également l'existence d'une classe supplémentaire sur les colonnes que nous supposerons être la classe 0 de telle sorte que si une colonne $j$ appartient à la classe 0 alors toutes les cases sont indépendantes et issues d'une même loi $\mathcal{B}(\lambda_j)$ :

$$X_{\cdot j} \sim \prod_{i=1}^{n} \mathcal{B}(\lambda_j),$$

avec $\boldsymbol{\lambda} = (\lambda_j)$ tel que $w_{j0} = 1$. Nous observons que le paramètre $\boldsymbol{\lambda}$ ne dépend pas de la structure en blocs, mais est spécifique à chaque colonne, ce qui offre une certaine flexibilité dans la modélisation du bruit. Par exemple, un petit $\lambda_j$ indiquera une colonne de bruit

prenant essentiellement la valeur 0, tandis qu'une valeur élevée indiquera une colonne comportant majoritairement des 1. Pour les colonnes pertinentes, nous supposons qu'elles sont issues d'une distribution de probabilité dont les paramètres sont dépendants de la structure en blocs, ce qui revient au modèle des blocs latents décrit précédemment.

Enfin, comme pour ce dernier, nous supposons que l'affectation des labels $w_j$ de chaque colonne est indépendante des autres affectations et suit une loi multinomiale $\mathcal{M}(1; 1-\phi, \phi\tau_1, \ldots, \phi\tau_m)$ où $\phi \in ]0;1[$ et toujours $\sum_{\ell=1}^{m} \tau_\ell = 1$. Le paramètre $\phi$ permet ainsi d'estimer la proportion de colonnes importantes à la classification des lignes. Finalement, nous pouvons définir le modèle des blocs latent avec une classe de bruit postulant qu'une matrice de données est issue du processus génératif suivant :

- Tirer $\mathbf{Z}$ selon $\prod_{i=1}^{n} \mathcal{M}(1; \boldsymbol{\pi})$ dans $\{1, \ldots, g\}$.
- Tirer $\mathbf{W}$ selon $\prod_{j=1}^{d} \mathcal{M}(1; (1-\phi), \phi\boldsymbol{\tau})$ dans $\{0, \ldots, m\}$.
- Tirer $\mathbf{X}$ avec pour chaque $j \in \{1, \ldots, d\}$ :
  - si $w_{j0} = 1$, $\mathbf{X}_{\cdot j} \sim \mathcal{B}(\lambda_j)^n$,
  - sinon $\mathbf{X}_{\cdot j} \sim \prod_{i=1}^{n} \mathcal{B}(\alpha_{z_i w_j})$.

La densité associée à ce modèle s'écrit alors

$$
\begin{aligned}
p(\mathbf{x}; \theta) &= \sum_{\mathbf{z},\mathbf{w}} p(\mathbf{x}|\mathbf{z}, \mathbf{w}; \theta) p(\mathbf{z}, \mathbf{w}; \theta) \\
&= \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \times (1-\phi)^{w_{+0}} \phi^{(d-w_{+0})} \prod_{j,\ell} \tau_\ell^{w_{j\ell}} \times \prod_{i,j} f(x_{ij}, \lambda_j)^{w_{j0}} \times \prod_{i,j,k,\ell} f(x_{ij}, \alpha_{k\ell})^{z_{ik} w_{j\ell}},
\end{aligned}
$$

où $f(x_{ij}, \alpha_{k\ell}) = \alpha_{k\ell}^{x_{ij}} (1-\alpha_{k\ell})^{(1-x_{ij})}$, $f(x_{ij}, \lambda_j) = \lambda_j^{x_{ij}} (1-\lambda_j)^{(1-x_{ij})}$ et $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ est l'ensemble des paramètres du modèle.

## 2.3 Point de vue bayésien

Comme suggéré par Keribin et al. [4], nous proposons une adaptation bayésienne. Pour ce faire, nous avons choisi les lois a priori suivantes :

$$
\boldsymbol{\pi} \sim \mathcal{D}ir(a, \ldots, a),\ \boldsymbol{\tau} \sim \mathcal{D}ir(a, \ldots, a),\ \boldsymbol{\alpha} \sim \prod_{k=1}^{g} \prod_{\ell=1}^{m} \mathcal{B}e(b, b),\ \phi \sim \mathcal{B}e(c_1, c_2) \text{ et } \lambda_j \sim \mathcal{B}e(e_1, e_2)
$$

où $\mathcal{D}ir$ est la loi de Dirichlet et $\mathcal{B}e$ la loi Bêta (voir la figure 1 pour une représentation schématique). Pour les paramètres propres au modèle des blocs latents, nous avons choisi de garder les mêmes lois a priori équilibrées proposées par Keribin et al. [4]. Toutefois, pour les nouveaux paramètres, nous laissons la possibilité d'une information a priori, en particulier, nous pourrions supposer que :

- la proportion de colonnes intéressantes représente plus de la moitié ($1 < c_2 < c_1$) ou moins de la moitié ($c_2 > c_1 > 1$),

— les colonnes de bruits aient des proportions de 0 et de 1 proches ($e_1 = e_2 > 1$) ou alors qu'elles contiennent tantôt quasiment que des 0 et tantôt quasiment que des 1 ($e_1 = e_2 < 1$).

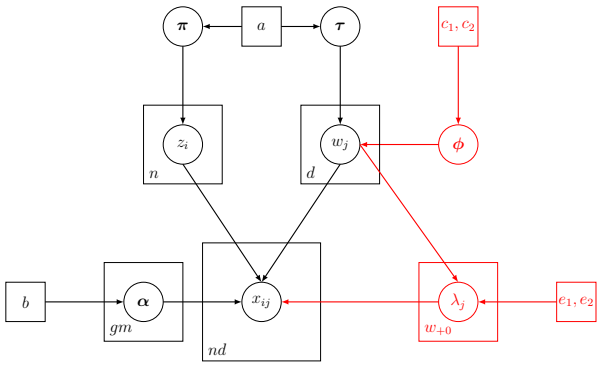


FIGURE 1 – Graphe bayésien du modèle des blocs latents avec une classe de bruit : en rouge sont représentées les parties rajoutées par rapport au modèle des blocs latents.

# 3 Estimations et premiers résultats

Pour l'estimation du nombre de classes et des paramètres, nous avons choisi de suivre la procédure proposée par Keribin et al. [4] utilisant l'échantillonneur de Gibbs, qui simule la loi a posteriori des paramètres, couplé avec l'algorithme *V-Bayes*, qui utilise une approximation variationnelle, sur une grille de couples $(g, m)$ de classes et en sélectionnant le couple maximisant le critère *Integrated Completed Likelihood* (voir [1]).

Cette méthode a été appliquée à des données génétiques, appelées les micro-satellites. Un micro-satellite est une séquence ADN formée par une répétition continue d'unités, généralement composée de 1 à 4 nucléotides. La longueur de ces séquences (i.e., le nombre de répétitions) varie selon les espèces, mais également d'un individu à un autre et même d'un allèle à un autre. Cependant, l'emplacement de ces séquences dans le génome est relativement similaire pour les espèces phylogénétiquement proches. Dans ce contexte, nous proposons d'appliquer notre modèle sur des données micro-satellites relevées chez différents individus à travers le monde (voir [5]). Notre modèle présente plusieurs avantages pour traiter ce type de données. D'une part les données génétiques sont généralement des données extrêmement bruitées, et l'approche classique est de filtrer en pré-traitement, les informations pertinentes avant d'effectuer des analyses, ce qui par conséquent ne tient pas compte de la structure des données. D'autre part, les approches de classification croisée sont particulièrement efficaces pour ce type de données car elles permettent d'identifier des groupes d'individus présentant des caractéristiques communes à un niveau local, ce qui est généralement plus pertinent en grande dimension.

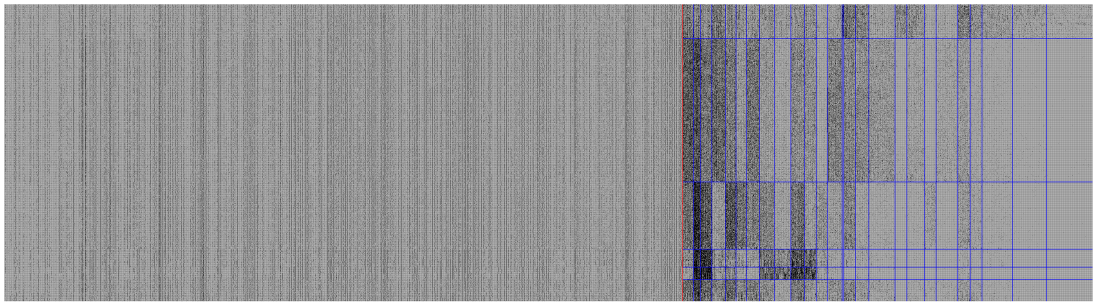

FIGURE 2 – Représentation de la matrice après réorganisation par les classes estimées : la droite rouge sépare les colonnes considérées comme du bruit (à gauche) de la partie dépendant du modèle des blocs latents dont les classes sont symbolisées par des traits bleus (à droite).

Sur la figure 2 est représentée la matrice réorganisée ; nous observons qu'une large partie des colonnes a été considérée comme du bruit. Parmi les classes en lignes, les individus sont essentiellement séparés suivant leurs continents d'origine.

# 4 Résumé

Dans cet exposé, nous développerons certains des résultats obtenus sur ce modèle comme l'identifiabilité et la consistance de l'estimateur du maximum de vraisemblance lorsque les $\lambda_j$ sont identiques ou sous d'autres conditions techniques (adaptés des résultats de Keribin et al. [4] et de Brault et al. [2]). Nous présenterons également les estimations obtenues sur des données simulées et réelles.

# Bibliographie

[1] Biernacki, C., Celeux, G. et Govaert, G. (2000). *Assessing a mixture model for clustering with the integrated completed likelihood.* IEEE transactions on pattern analysis and machine intelligence, 22(7), 719-725.
[2] Brault, V., Keribin, C. et Mariadassou, M. (2017). *Consistency and asymptotic normality of latent blocks model estimators.* arXiv preprint arXiv :1704.06629.
[3] Govaert, G., et Nadif, M. (2003). *Clustering with block mixture models.* Pattern Recognition, 36(2), 463-473.
[4] Keribin, C., Brault, V., Celeux, G. et Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6), 1201-1216.
[5] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. et Feldman, M. W. (2002). *Genetic structure of human populations.* science, 298(5602), 2381-2385.